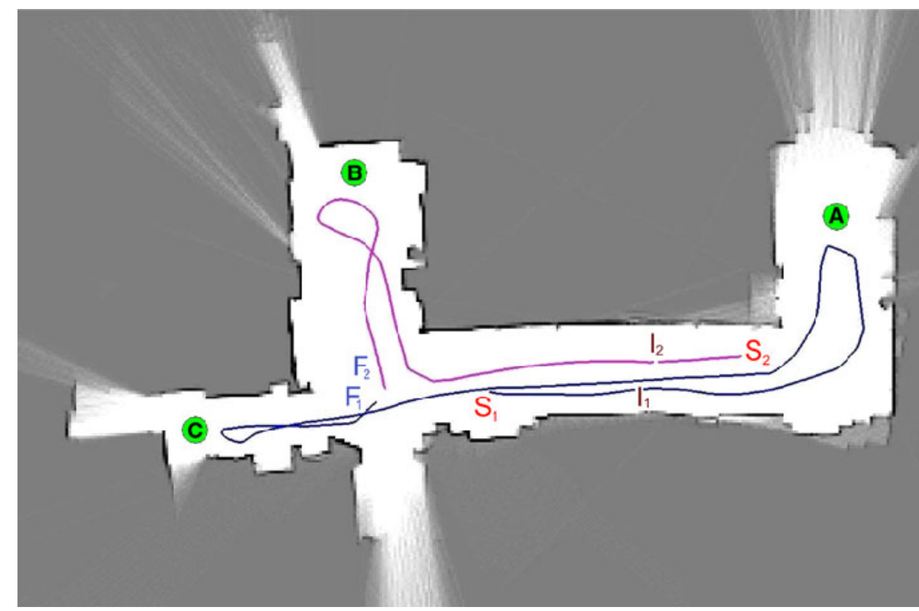
Partially observable systems

Partially observable systems

- Key assumption so far: Markov property
- Real-world is non-Markov / partially observable (PO)
    - Or you wouldn't need *memory*
- Examples in ML

Alan Mathison Turing OBE FRS (/ˈtjʊərɪŋ/; 23 June 1912 – 7 June 1954) was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist.[2] Turing was highly influential in the development of theoretical computer science, providing a formalisation of the concepts of algorithm and computation with the

text modeling (last word cannot predict what's next; need to capture long-term dependencies)



SLAM in robotics ("this place looks familiar; *did I return to the same location*?")

"perceptual aliasing"

Prev. frame   Next frame

video prediction

Models of PO systems

- Observation space O
- Actions space A (omitted in most discussions)
- System starts from initial configuration, and outputs sequences $o_1\, o_2\, o_3 \ldots$ with randomness
- Markov systems is a special case:

$$\Pr[o_{t+1:t+k} \mid o_{1:t}] = \Pr[o_{t+1:t+k} \mid o_t]$$

$$\text{or, } \; o_{t+1:t+k} \perp o_{1:t} \mid o_t \quad \text{(treated as r.v.'s)}$$

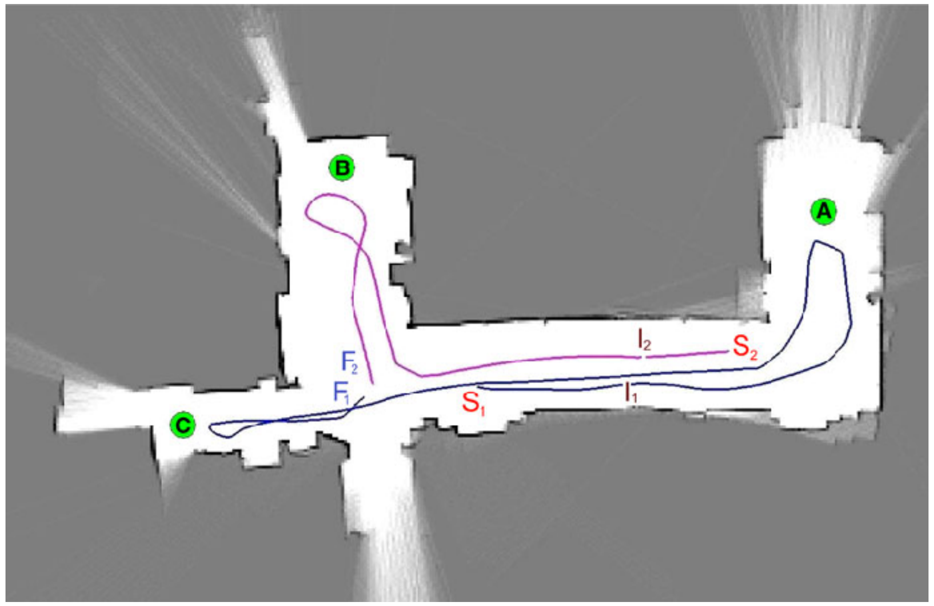
  - In words, last observation is *sufficient statistics of history* for predicting future observations
- How restrictive is Markov assumption?

Complexity of Markov vs non-Markov systems

- For a Markov chain, the complexity is measured by the number of states (i.e., number of observations)
  - System fully specified by the transition matrix $T(o' \mid o)$
  - # model parameters = $|O|^2$
- Without Markov assumption?
  - System fully specified by $\Pr[o' \mid h]$ for any history h (short for $o_{1:t}$)
  - Probabilities for different histories can be set completely independently— with horizon L, order $|O|^L$ free parameters!
  - Even with a finite and small observation space, fully general dynamical systems are intractable
  - Need structure…

Partially observable systems

- Example structure: small & finite *latent* state space
- “this place looks familiar; did I return to the same location?”
  - No structural assumption: you always visit a new location
  - With structural assumptions: the building only has this many rooms. You will be in one or another.

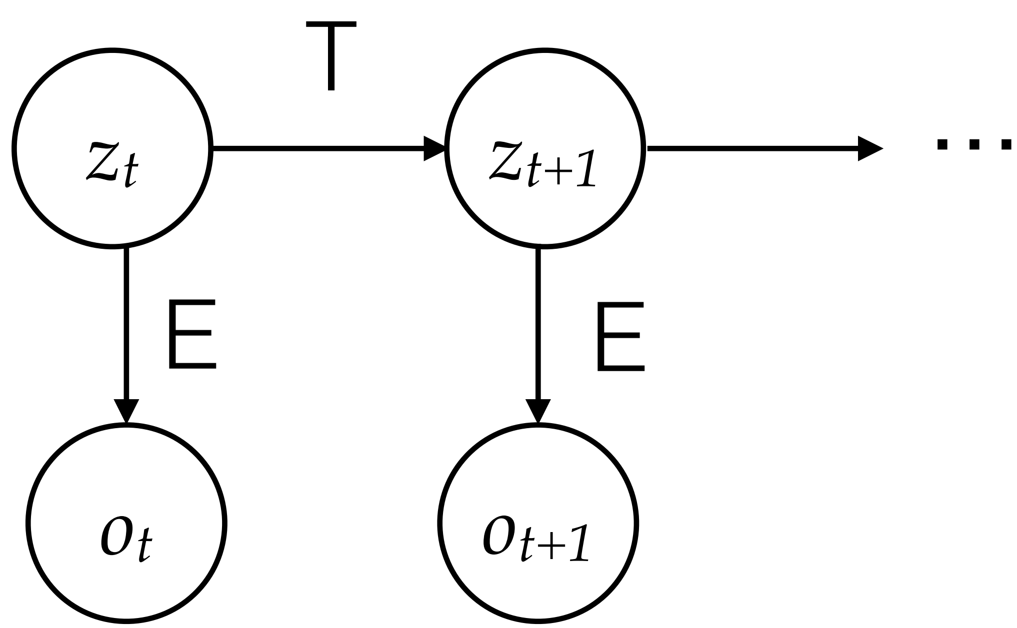


SLAM in robotics (“this scene looks familiar; *did I return to the same location*?”)

Latent Models of PO systems

- Observation space O
  - SLAM example: current sensory inputs
- Action space A (again will be ignored in most places)
- Latent/hidden state space Z
  - SLAM example: true location
  - Implicit assumption: Z is "simple" (e.g., finite & small)
- Model parameters
  - Emission probability: $E(o \mid z)$
  - Transition probability: $T(z' \mid z)$ (controlled case: $T(z' \mid z, a)$
  - Sometimes, also the initial distribution: $d_0(z)$
- Markov chain is special case: identity emission

Graphical representation



Myth 1 about HMMs/POMDPs

- PO can stem from noisy sensors, which compresses/loses information from “world state”
- Noisier sensors = more PO?
- Mathematically, if we fix the underlying MDP and vary the emission function, an emission that loses more information gives a more PO process?
- Wrong: If emission discards all information, the process becomes Markov!

Myth 2 about HMMs/POMDPs

- When the problem is non-Markov, people will say "oh it's a POMDP"
- …which assumes POMDP is fully general?
- Not really: there are systems that can be succinctly represented but require infinitely many hidden states to be represented as a POMDP/HMM
- Again, the most general way to specify a PO system is just $\Pr[o_{t+1}{=}o' \mid o_{1:t}]$, or $\Pr[o' \mid h\ ]$ for short (h for history)
    - *any* (possibly PO) environment is equivalent to an MDP whose state is the history in the original environment

Major challenge in PO systems: *state* representation

- Examples
  - Text prediction: how to *compactly summarize* the sentence so far to predict future words? (that's what you are computing as the hidden vector in an LSTM)
  - SLAM: how to map history of sensor readings to physical locations (or a *belief* about it)
- What does state mean in the PO setting?

Definition: **State** is a **function of history**, ϕ, that is a **sufficient statistics** for **predicting future**. That is, for all $e := o_{t+1:t+k}$ and $h := o_{1:t}$,

$$\Pr[e \mid h] = \Pr[e \mid \phi(h)]$$

Computing a compact state given the model

- Suppose we know the HMM model $E(o \mid z)$, $T(z' \mid z)$, $d_0(z)$
- How to compactly summarize any history $o_{1:\tau}$?
- Belief state: $\phi(h) = [\mathbb{P}[z_{t+1} = z \mid h]]_{z \in Z} \in \mathbb{R}^{|Z|}$ where $t := |h|$
  - belief state *is state*
- Computing belief state
  - Initialization: $\phi(\emptyset) = d_0$ ($\emptyset$ is empty history)
  - Update using Bayes rule: if we know $\phi(h)$, then we can compute $\phi(ho)$ as (ho is the concatenation of h and o)

$$\mathbb{P}[z_{t+2} = z' \mid ho] = \frac{\mathbb{P}[z_{t+2} = z', o_{\tau+1} = o \mid h]}{\mathbb{P}[o_{t+1} = o \mid h]}$$

  - Enumerator:

$$\mathbb{P}[z_{t+2} = z', o_{t+1} = o \mid h] = \sum_{z \in Z} \mathbb{P}[z_{t+2} = z', o_{t+1} = o \mid h, z_{t+1} = z] \, \mathbb{P}[z_{t+1} = z \mid h]$$

$$= \sum_{z \in Z} T(z' \mid z) E(o \mid z) \, \mathbb{P}[z_{t+1} = z \mid h]$$

Computing a compact state given the model

- Matrix form: Let T be the |Z|x|Z| transition matrix, and E_o be a |Z|x|Z| diagonal matrix whose z-th diagonal entry is $E(o \mid z)$
- $\phi(ho) \propto T E_o \phi(h)$
- The matrix form is also useful for making predictions, e.g., $\mathbb{P}[o_{t+1:t+k} \mid h] = \mathbf{1}^\top T E_{o_{t+k}} T E_{o_{t+k-1}} \cdots T E_{o_{t+2}} T E_{o_{t+1}} \phi(h)$
- The controlled case:
  - define T_a as the |Z|x|Z| matrix, whose (z',z)-th entry is $T(z' \mid z, a)$
  - To compute belief state and make predictions: replace $T E_o$ above by $T_a E_o$
  - e.g., $\mathbb{P}[o_{t+1:t+k} \mid h, a_{t+1:t+k}] = \mathbf{1}^\top T E_{o_{t+k}} T_{a_{t+k}} E_{o_{t+k-1}} \cdots T_{a_{t+2}} E_{o_{t+2}} T_{a_{t+1}} E_{o_{t+1}} \phi(h)$
  - meaning of LHS: at time t, if the history is h, and we will take actions $a_{t+1:t+k}$ for the next k steps, what is the probability that we observe $o_{t+1:t+k}$?

State!

- Trivial function that is state?
  - History itself (identity map): $\phi(h) = h$
  - There is another one: $\{\Pr[e \mid h]\}_{e \in E}$ where *E* is the (infinite) set of all future events
- For HMMs/POMDPs, belief state, $(\Pr[z_\tau = z \mid h])_{z \in Z}$, *is state*
- To an old-school RL person, be careful when you say "state" without a modifier…
- Things that are not states and people call "state"
  - Observation: e.g., Atari game frame
  - Hidden state ("World state") : Why?
  - Agent state: can be approximately a state

Policy optimization in a POMDP

- Consider a POMDP that is specified by:
  - Emission probability: $E(o \mid z)$
  - Transition probability: $T(z' \mid z, a)$
  - Initial distribution of hidden state: $d_0(z)$
  - Reward function: $R(z, a)$
  - And some notion of horizon (e.g., a finite horizon of H)

  MDP
- We'd like to link to familiar concepts in MDPs…
  - Any POMDP is equivalent to an MDP where history of observations & actions is treated as state
  - Value functions & optimal policies immediately well-defined!
  - Conceptually useful but practically not—the number of states is exponentially in H
  - (Actually, planning in POMDP is hard anyway (PSPACE-complete))

Policy optimization in a POMDP

- We know that POMDP is also equivalent to another MDP…
  - whose state is the belief state: $b(h) \in \mathbb{R}^{|Z|}$
  - Then we get a continuous MDP whose state space is $\mathbb{R}^{|Z|}$
- How to define the parameter of this MDP?
  - Transition: in any (belief) state $b \in \mathbb{R}^{|Z|}$, if we take action a, then the distribution of next (belief) state b' follows the below generative process:

$$z \sim b, \;\; z' \sim T(\cdot \,|\, z, a), \;\; o' \sim E(\cdot \,|\, z'), \;\; b' = \phi(ho')$$

  - Similarly, $R(b, a) = \sum_{z \in Z} b(z) R(z, a)$
- Compared to history-based MDP (exponentially many discrete states), the belief-state MDP has a continuous state space…
  - but it is more structured! If two belief vectors are close, the value functions are also close
  - can approximate by e.g., discretization

Policy optimization in a POMDP

- There is more than smoothness…
- Given a fixed deterministic policy π (that maps belief states to actions), its value function V^π is linear in b:
  $V^\pi(b) = \langle b, [V^\pi(b,z)]_{z \in Z} \rangle$; $[V^\pi(b,z)]_{z \in Z}$ is often called an α-vector
- Implies that V^* is piece-wise linear in b, since there are only finitely many policies (assuming finite observation space and finite horizon)
- Sometimes a policy is dominated by other policies and can be pruned
- A popular approach: dynamic programming from bottom and prune α-vectors before applying Bellman eq

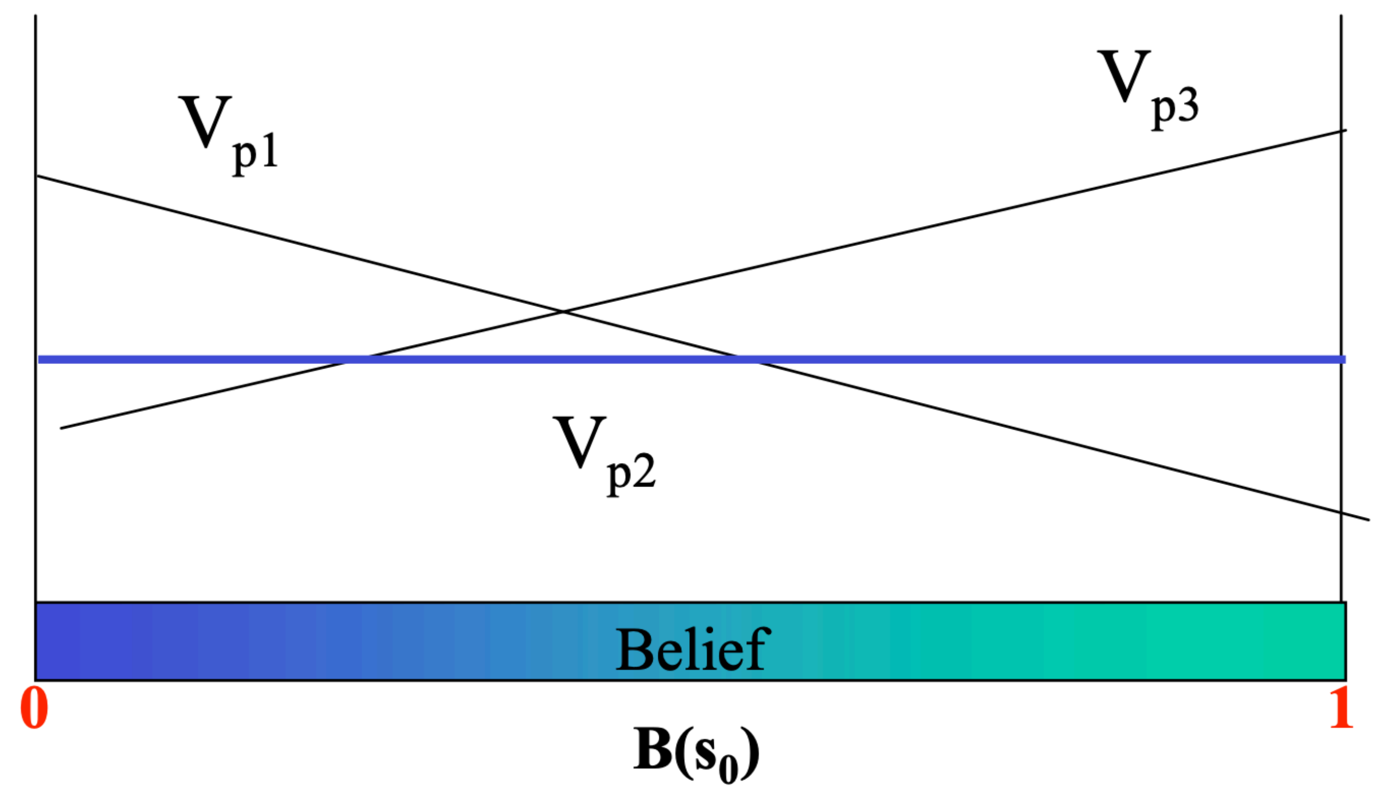


Fig credit: https://www.techfak.uni-bielefeld.de/~skopp/Lehre/STdKI_SS10/POMDP_tutorial.pdf

Learning partially observable systems

- So far we've been talking about how to compute belief state and optimal policy given the HMM/POMDP model
- How to learn such a model from data?
- Standard approach: EM (Expectation-Maximization)
  - Consider HMM. Say our data are sequences of observations in the form of $o_{1:\tau}$
  - E-step: pretend that the current estimated model were true, calculate the posterior over hidden states (given data)
  - M-step: pretend that the posterior were true, update the estimated model to be the maximum likelihood model given data (observation seq) + posterior over hidden states
  - Repeat
- Alternative approach: spectral learning (Method of Moments)